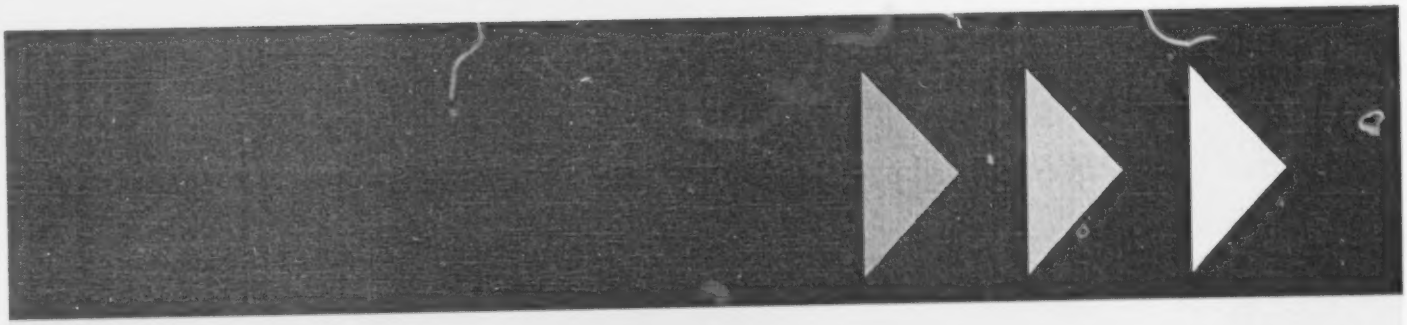
Treasury Board of Canada Secretariat
Secrétariat du Conseil du Trésor du Canada

Review of the Quality of Evaluations Across Departments and Agencies



Canada

105-03644/

Acknowledgements

A Working Group was established to provide input and context for the Review. We would like to thank the following participants:

- CFIA - Theresa Iuliano
- Health Canada - Walter Zubrycky
- Western Economic Diversification Canada - Kathy Locke
- Correctional Services Canada - Christa Gillis
- Transport Canada - Unnati Vasavada
- HRDC - Serge Bertrand
- AAFC - Eric Seraphim
- DFAIT - Stephen Kester

They provided feedback on the terms of reference for the study, suggestions regarding the Review Template and comments on the draft report.

We are most grateful to Glenn Crone and Zeljka Spasojevic of the Centre of Excellence for Evaluation, Treasury Board Secretariat, for their ongoing support.

Members of the Working Group worked in collaboration with Shelley Borys, Michael Callahan, Janice Remai of EKOS research Associates, Inc.

Table of Contents

Executive Summary 1
Introduction 1
Purpose 1
Methodology 1
Findings 2
Conclusions and Recommendations 4

1. Introduction 5
   1.1 TBS Evaluation Policy 5
   1.2 Centre of Excellence for Evaluation 6
   1.3 Organization of the Report 7

2. Methodology 7
   2.1 Design of Guide for the Review of Evaluations 7
   2.2 Sampling 9
   2.3 Review of Evaluation Reports 10
   2.4 Analysis 11

3. Findings 12
   3.1 Quality of Federal Evaluations: Overview and Highlights 12
   3.2 Detailed Findings 16
   3.3 Strengths and Weaknesses of Federal Evaluations 32
   3.4 Variations in Quality by Organizational Characteristics and Report Date 35

4. Conclusions and Recommendations 40
   4.1 Conclusions 40
   4.2 Recommendations 41

Appendix A Review Template 43

Appendix B Distribution of Reviewed Reports by Department/Agency 63

Executive Summary

Introduction

Evaluation supports the Government of Canada's aim to becoming a learning organization. It does this by helping senior executives, program managers and policy makers discover whether or not their initiatives work and are meeting objectives, whether or not there is a continued need for their initiatives, and how their initiatives can be better designed and delivered to meet objectives in a cost-effective manner. The quality of evaluation reports is fundamental if the evaluation function is to deliver upon these information needs.

Purpose

In 2001, TBS created the Centre of Excellence for Evaluation (CEE) and established a new Evaluation Policy to strengthen the evaluation function and the quality of reporting. A key objective of this report is to address whether the quality of reports is acceptable and whether there has been an improvement in quality. An important aspect of this work is to promote quality evaluation reports. This review represents one piece of CEE's overall strategy to monitor and strengthen the quality of reporting. Other activities include: best practice research; an annual survey of the health of departmental and small agency evaluation units; individual meetings; ongoing review of evaluations, RMAFS, departmental evaluation plans; and, an annual report documenting evaluation findings and how they contribute to strengthening accountability and the government's Expenditure Review exercise.

Methodology

A number of sources were to develop the criteria used for this review including the Guide for the Review of Evaluation Reports", prepared by the Centre of Excellence for Evaluation, TBS, January 2004 and excerpts from the OAG 1993 Report on Program Evaluation ("Criticisms re Evaluation Reports"). A reference group of department and agency evaluation units was also consulted. The template used for the review is presented in Appendix A. 1

1. The original intention was to use a stratified sample of evaluation reports according to key variables of interest. As it turned out, the population of reports to be considered for this review consisted of only those evaluation reports that have been submitted to TBS. Although departments are requested to submit all completed evaluation reports, it appears that they do not do this reliably. Based on the capacity check research conducted by CEE two years ago, it appears that approximately 250 evaluations are completed each year, which should have resulted in 500 reports being available for review. However, only 214 completed reports have been submitted to TBS over the last two years (the years of interest in this review). In addition, many evaluation records are on file (e.g., web-links, reviews), do not meet the definition of "complete, hard-copy of an evaluation available for the purpose of the review". Given the absence of the full population from which to sample, it is difficult to assess the degree to which the pool of reviewed reports is biased in any way. The distribution of reviewed reports by department/agency is presented in Appendix B.

Findings

The findings of this review indicate that most federal evaluation reports are acceptable in quality, though almost one-quarter of the evaluations (23 per cent) were rated as inadequate overall. No clear and consistent variations in quality were observed for federal organizations of different sizes and for departments versus agencies. A comparison of reports completed pre- versus post-April 2002 indicates, however, that quality has improved on a number of criteria in the more recent evaluations. For example, this includes: addressing cost-effectiveness issues; methodological rigour; identifying alternatives; presentation of evidence-based findings; and, formal recommendations. This increase in quality over time suggests that TBS's efforts to improve the quality of evaluation may be having a positive impact (i.e., allowing one year, until April 2002, for the Policy to be fully understood by departments/agencies and for the new Centre for Excellence in Evaluation to begin operating). Still, there is a pressing need for further improvement as indicated by the findings noted below.

Key ***strengths*** of the evaluations examined in this review include:

- a comprehensive description of the program/initiative under review including its resources, beneficiaries and stakeholders;
- a clear statement of the evaluation objectives;
- the use of multiple lines of evidence in the methodology;
- a strong presentation of findings, in particular, on relevance and delivery/implementation issues;
- formal recommendations or suggested improvements that flow logically from the findings and conclusions; and
- reports that are well-written and well-organized.

On the other hand, some weaknesses of evaluations/reports are:

- only six in ten evaluation reports indicated the timing and significance of the evaluation;
- most of the reports only listed (two-thirds) and very few discussed (about one-quarter) the evaluation issues;
- superficial coverage of cost-effectiveness issues;
- many reports lacked a full description of the key methodological details. While just over half of reports described the methodology, four in ten only listed a few details and only one-quarter referenced a technical document;

- there is little incorporation of data from a performance measurement system;
- only a minority of the evaluation designs included features to optimize the rigour of the research such as a comparison group (13 per cent), baseline measures (14 per cent) or a comparison to norms, literature or some other benchmark (22 per cent). Only 26 percent used interviews with independent key informants with no stake in the program;
- only about four in ten evaluation reports included a statement of the limitations or constraints of the evaluation;
- only about one-third of evaluations presented findings on whether the program duplicates or works at cross purposes with other programs/initiatives;
- only one-quarter of the evaluations discussed unintended outcomes (25 per cent) or addressed incremental impacts (26 per cent);
- only 26 per cent of evaluations provided findings on alternative, potentially more cost-effective approaches, though coverage of this issue has increased in more recent reports (31 per cent post-April 2002 versus 16 per cent pre-April 2002);
- almost one-quarter of evaluations (24 per cent) were rated as inadequate in their provision of objective, evidence-based conclusions related to relevance, success and/or cost-effectiveness;
- among the reports with recommendations, only 26 per cent identified alternative scenarios ; and,
- less than half of the evaluation reports included a management response (48 per cent) or action plan (33 per cent).
- 25% of reports with recommendations included a recommendation related to overall funding, and in all of these cases, the recommendation was to increase funding.
- No reports presented evidence indicating that a program was not relevant or not needed.

Conclusions and Recommendations

On balance, most evaluations that were assessed in this review are of reasonable quality. The majority received an overall rating of adequate (45 per cent) or more than adequate (32 per cent). Still, a considerable proportion of the evaluations (23 per cent) were rated as inadequate and this finding warrants attention. To this end, the report recommends that the TBS Centre of Excellence for Evaluation:

1. Encourage evaluation divisions in federal departments and agencies to strengthen their evaluation reports by addressing the major weaknesses identified in this review;

2. Refine Treasury Board guidelines/criteria for the expected features of (1) evaluation methodologies and (2) evaluation reports and disseminate them;

3. Continue to implement a rigorous approach to monitoring the quality of evaluations, and use this as a basis for the development of individual report cards on the quality and overall health of the evaluation function by department and small agency; and,

4. Identify measures, including an incentive structure and standards, to ensure that departments and agencies submit completed evaluations and reviews in a responsible, reasonable manner. Departments and agencies adherence to such standards should be made a public record.

1. Introduction

Evaluation supports the Government of Canada's aim to becoming a learning organization. It does this by helping program managers and policy makers to discover whether or not their initiatives work and are meeting objectives, whether or not there is a continued need for their initiatives, and how their initiatives can be better designed and delivered to meet objectives in a cost-effective manner. The Treasury Board Secretariat (TBS) introduced the Evaluation Policy (the Policy) in April, 2001, to clarify the important role of evaluation it its management framework.

The Centre of Excellence for Evaluation (CEE) was established in 2001 to assist with the implementation of the new Evaluation Policy, as well as to monitor the Policy's success. The CEE, in monitoring evaluation practices across federal departments and agencies, determined that there was a need to review the level of quality of these departments' and agencies' evaluations, with a view to identifying strengths and weaknesses of evaluation practices as well as appropriate responses. This document presents the Draft Final Report for this review of federal government evaluations.

1.1 TBS Evaluation Policy

Given the environment of renewal in the federal government, the importance of evaluation has risen considerably, but capacity to undertake it has not[2]. Resources, human and otherwise, devoted to evaluation have diminished steadily since the early 1990s. Furthermore, the current Evaluation Policy has increased the scope of work necessary under evaluation.

The TBS Evaluation Policy was last revised on April 1, 2001, and supports an "ongoing commitment to continuous management improvement and accountability," as stated by Minister Robillard in a February 14th , 2003 Press Release[3]. In the current Evaluation Policy, evaluation has a key role in supporting managing for results in the Public Service. The Policy rests on three principles: achieving and reporting on results is the responsibility of public service managers; rigorous, objective evaluation is an important tool in managing for results; and departments and agencies, with the support of the TBS are responsible for ensuring the evaluations are rigorous. The stated objective of the Policy is "to ensure that the government has timely, strategically focused, objective and evidence-based information on the performance of its policies, programs and initiatives to produce better results for Canadians." Its key requirements are as follows:

- Establishment of an appropriate evaluation capacity, including senior management;

---

2 Treasury Board of Canada Secretariat (September 2003). *Evaluation Policy: Results-Based Management and Accountability Framework (RMAF)*. page 4. Online: http://www.tbs-sct.gc.ca/eval/tools_outils/polrmaf-polcgrr-PR_e.asp?printable=True

3 Evaluation Policy: Results-Based Management and Accountability Framework (RMAF), opt. cit.

- Encompassing a wider scope, including policies, programs and initiatives, as well as those delivered through partnership mechanisms (e.g., inter-departmental and inter-governmental);
- Placing greater emphasis on performance monitoring and early results through:
    - Results-based Management and Accountability Frameworks (RMAFs) for new or renewed policies, programs and initiatives;
    - ongoing performance monitoring and measurement activities;
    - addressing issues related to early implementation and administration; and
    - addressing issues related to relevance, results and cost-effectiveness;
- Development of strategic evaluation plans;
- Integrating evaluation with management and strategic decision-making; and
- Implementing simplified and consolidated Standards of Practice.

1.2 Centre of Excellence for Evaluation

The CEE was established concurrent with the Evaluation Policy to provide leadership and aid in the implementation of the Policy. The current review of the quality of evaluation will support the CEE's mandate of monitoring and reporting on the state of evaluation capacity across the federal government. The CEE has been designed to serve the following key functions:

- to serve as a focus for leadership in federal government evaluation;
- to forge ahead on shared challenges such as devising a human resources framework for long-term recruiting, training and development needs; and
- to support capacity building, improving practices, and a stronger federal government evaluation community.

To these ends, the CEE carries out activities such as: policy implementation; monitoring; capacity building; strategic advice; and communications and networking.

1.3 Organization of the Report

This document contains the results of the "Review of the Quality of Evaluations across Departments and Agencies." Our methodology is presented in the next chapter. Findings are presented in Chapter 3 and conclusions from this work are in Chapter 4.

2. Methodology

This chapter describes the methodological approach to this project. The description is broken down into four sections: design of the review guide; sampling; review of evaluation reports; and a note on analysis.

2.1 Design of Guide for the Review of Evaluations

There were several sources which were assessed in the development of criteria for the purpose of this review. In casting about for possible indicators of quality for which information would be collected in this review, we initially turned to the Results-based Management Framework (RMAF) for the Treasury Board Secretariat's Evaluation Policy. An examination of the RMAF revealed that the review will specifically help to address the group of questions listed under Section E of Progress/Success Issues, namely: "Is the evaluation function of departments producing timely and effective insight, integrated with department decision making?" contributing to the Policy's immediate expected outcomes, which are *evidence-based reporting and timely credible reporting*. However, given the scope of this project, the timeliness of the reports cannot be assessed. Moreover, only evaluation reports completed *since* the Policy was implemented were reviewed here, so there is no baseline measure of quality against which to compare the results of the current review.

There are several documents addressing the issue of quality criteria which were consulted during the design of this work. These documents include:

- "Guide for the Review of Evaluation Reports", prepared by the Centre of Excellence for Evaluation, TBS, January 2004;
- "Checklist Form for Internal Control of Evaluation Study: Deliverables/Reports, Processes and Contractors' Work", prepared by Program Evaluation, HRDC, September 2003;
- "Health Canada Evaluation Report Assessment Guide", prepared by the Departmental Program Evaluation Division, Health Canada, April 2003;
- a framework for assessing the quality of evaluations, prepared by an external consultant for use by the Office of the Auditor General (but not implemented); and
- excerpts from the OAG 1993 Report on Program Evaluation ("Criticisms re Evaluation Reports"), prepared by CEE.

The core research questions centred around the following: Is the quality of reports acceptable and has there been an improvement in the quality? Note that, with a mere review of evaluation reports, we are not able to determine if there has been an improvement in the quality of the reports. Such information can be collected only through comparisons with pre-Policy evaluations and interviews with officials. However, based on a review of the Evaluation Policy (including Appendix B of the Policy), its RMAF, and the materials referred to above, potential indicators that were identified to measure the quality of evaluation reports include the following characteristics:

- are clearly written, are concise and use simple language;
- clearly describe the program, policy or initiative being evaluated, including its objectives, outputs, expected outcomes, reach, and resources;
- have an assessment of the results achieved by the policy, program or initiative;
- have a description of the evaluation, including its timing; the methodology; the evaluation objectives and issues; and how the evaluation fits into, and is important to, the overall operations of the department or agency;
- expose the limits of the evaluation, in terms of context, scope, methods and conclusions;
- have an appropriate methodology (e.g., multiple lines of evidence);
- have conclusions that clearly address the main evaluation issues of relevance, success/impacts, and cost-effectiveness (depending on the type of evaluation - formative or summative);
- include only information necessary to understand findings, conclusions and recommendations;
- present evidence-based and credible findings, for example:
    - evidence gathered in surveys of a representative group of participants, and compared to a comparable group of non-respondents;
    - evidence derived from comparisons to baseline measures from the performance measurement system; and
    - qualitative evidence gathered from key informants who do not have a stake in the respective program or who are truly knowledgeable in the area of question;
- have conclusions and recommendations flowing logically from evaluation findings;
- have clear, attainable recommendations indicating actions to be taken and time frame; and

- provide analysis and explanation of exposure to risk of problems identified and in respect to recommendations made.

Based on our analysis of all reference material indicated above, a draft template was prepared for the review. Following the development of a draft instrument containing proposed criteria and review of this with the project authority, we met with the CEE's working group (which represented eight different federal departments) to discuss the criteria and the scope of the review. Comments received at that time were taken into account in revisions to the review template. The final template used for the review is presented in Appendix A.

2.2 Sampling

We had proposed that the sample of evaluation reports would be selected from a database compiled by the CEE of reports on evaluations conducted since the inception of the Evaluation Policy, i.e., the 2001/02 fiscal year. The "population" of reports would be stratified according to certain key variables of interest. Titles of reports would be selected in terms proportional to population characteristics, or in sufficient numbers to ensure representation from all key sub-groups.

To the degree that stratification was possible and/or desired, there were a number of potential sample stratification/selection variables, for example: the type of evaluation, formative or summative; the size and type of department or agency; the year of the evaluation (as the quality of evaluations may be expected to rise over time, as the Policy takes hold and evaluators and CEE officials become increasingly familiar with it).

As it turned out, the population of reports to be considered for this review consisted of only those evaluation reports[4] which have been submitted to TBS. Although departments are requested to submit all completed evaluation reports, it appears that they do not do this reliably. Based on the capacity check research conducted by CEE two years ago, it appears that approximately 250 evaluations are completed each year, which should have resulted in 500 reports being available for review. However, only 214 complete reports have been submitted to TBS over the last two years (the years of interest in this review). In addition, other evaluation records are on file (e.g., web-links, reviews), but did not meet the definition of "complete, hard-copy of an evaluation available for the purpose of the review".

Given the limited timeframe available for this review, it was not possible to obtain the full set of evaluation reports from individual departments. Further, it is difficult to determine the impact on

4 The reports in the population and our sample (n=115) included both mandatory and non-mandatory evaluation reports. Mandatory evaluations (i.e., those done to support a TB submission for program funding renewal) focus on specific issues (e.g., those specified in the RMAF) so TB has clear guidelines as to what should be addressed in these reports. In contrast, non-mandatory evaluations can have a narrower or broader focus, depending on their purpose.

the objectivity of the sample were we to have canvassed departments and agencies and ask them to submit reports for purposes of this review.

Thus, it is important to note that the review was conducted with this limited sample of evaluation reports which have been submitted to TBS and where the files are complete. Given the absence of the full population from which to sample, it is difficult to assess the degree to which the pool of reviewed reports is biased in any way.

In the process of locating reports for review, the full set of reports submitted after April 1, 2001 and available through TBS was accessed. Although the database indicated that there were more than 200 reports available for this exercise, many of the files were determined to not be appropriate for review. For example, some files contained only an executive summary of a report, or were reports on audits or special studies (e.g., to provide in-depth research on a topic that would feed into an evaluation, but not be an evaluation itself) or other types of review that did not constitute an evaluation.

The work plan had been to review a total of 110 reports. Ultimately, we had 122 reports which were available for review, and of these 115 were reviewed. Those that were not reviewed (n=7) were reports from departments that were already heavily represented in the sample. We attempted to limit the total number of reports reviewed for any one department, to ensure representation across the population of reports available. As it turned out, several departments had 10 to 12 reports which were reviewed (and these departments were also the ones with reports available but not reviewed).

Six reports in the sample had been prepared by the CEE. As it was inappropriate for us to review our own reports, analysts from TBS were trained in the application of the review template and then completed the reviews of five of the six of these (a sixth report was for a department that was already well-represented and thus, was not needed).

The distribution of reviewed reports by department/agency is presented in Appendix B.

2.3 Review of Evaluation Reports

An extensive pretest process involving all reviewers was undertaken, not only to test the review template, but also to ensure inter-rater reliability. A total of three reports were reviewed by each of the core team members. After the review and completion of the template for each report, the team met to thoroughly discuss the ratings each had assigned. Where there were discrepancies, subsequent discussion enabled clarification of the meaning of certain review aspects or ratings. As well, the template was revised to accommodate this additional clarification where possible. The revised template would then be used for the next pretest review. By the end of the third pretest review, inter-rater reliability (assessed qualitatively) was determined to be sufficiently high to proceed with independent reviews.

Following the pre-test and finalization of the review template, the full review of evaluations was conducted. Each evaluation report was assessed by only one reviewer. All reviewers were knowledgeable evaluators with considerable experience in the evaluation of federal programs. Each report review took an average of 2.5 hours to complete.

2.4 Analysis

Univariate and cross-tabulation analyses were run on the data from the reviews. Most of the criteria assessed in the reviews were rated on a five-point scale ranging from 1 ("poor") to 5 ("excellent"), with the mid-point 3 indicating "adequate". For the analyses, the scale ratings were collapsed into the three following categories: 1-2 ("inadequate"), 3 ("adequate") and 4-5 ("more than adequate"). Cross-tabulations based on size of the department/agency were then conducted. Three categories were developed: small (500 FTEs or less, n=18) 5; medium (501 to 4,600 FTEs, n=51); and large (more than 4,600 employees, n=46). In addition, cross-tabulations were run on the year of the report (up to March 2002, n=37, and April 2002 and beyond, n=78) and also on department (n=91) versus agency (n=24). The tables of results are presented in a Technical Appendix under separate cover.

A) Limitations

The quality of evaluations can be measured in different manners. In this review, we looked at the quality of the evaluations as reflected in the evaluation reports. It is important to note that another important dimension of the quality of evaluations, not examined in this review, is their usefulness as reflected in the degree of implementation of evaluation recommendations. CEE has indicated that it will examine this element of quality though other lines of evidence.

It is important to note that, as external reviewers of an evaluation report, we did not always have full information on potential limitations to any particular evaluation (e.g., budget restrictions, available timeframes, internal constraints) or the context (we did not interview evaluation or program managers). Thus it is possible that some reports may be considered weak in our review, although perhaps given the external limitations facing them or the context, they may in fact have been quite strong.

The CEE working group also suggested that the quality of evaluation reports assessed in this review may appear to be weak in some regards because departments were not aware of the assessment criteria at the time the evaluations were undertaken. In addition, the working group suggested that departments may have examined or addressed criteria in the assessment criteria used in this review but did not report on them in the evaluation report because they were reported elsewhere or because they considered them to be not relevant for the report.

5 Given the small number of reports from small organizations, findings related to this category should be treated with caution.

In addition, due to time and budgetary constraints on the present review (i.e., only 2.5 hours were available to review each report), it was determined with the client during the design phase that the review would be predominantly quantitative (i.e., closed items in the review template presented in Appendix A). Consequently, detailed qualitative information explaining the various ratings for each evaluation report was not collected.

3. Findings

3.1 Quality of Federal Evaluations: Overview and Highlights

The findings of this review indicate that most federal evaluation reports are acceptable in quality, though almost one-quarter of the evaluations (23 per cent) were rated as inadequate overall. No clear and consistent variations in quality were observed for federal organizations of different sizes and for departments versus agencies. A comparison of reports completed pre- versus post-April 2002 indicates, however, that quality has improved on a number of criteria in the more recent evaluations. This suggests that TBS's April 2001 Evaluation Policy may have had a positive impact (i.e., allowing one year, until April 2002, for the Policy to be fully understood by departments/agencies and for them to implement some improvements). Still, there is a need for further improvement as indicated by the weaknesses noted below.

The review reveals that federal evaluation reports have a number of strengths as well as limitations, though there is no clear pattern to these (i.e., a given section of the reports, such as the introduction/context, exhibits both strengths and weaknesses depending on the particular criterion assessed). Key strengths of the evaluations examined in this review include:

- a comprehensive description of the program/initiative under review including its resources, beneficiaries and stakeholders;
- a clear statement of the evaluation objectives;
- the use of multiple lines of evidence in the methodology;
- a strong presentation of findings, in particular, on relevance and delivery/implementation issues;
- formal recommendations or suggested improvements that flow logically from the findings and conclusions; and
- reports that are well-written and well-organized.

On the other hand, some weaknesses of evaluations/reports are:

- neglecting to present or reference the program logic model;

- inadequate discussion of the evaluation issues and failing to reference source documents sch as RMAFs or Evaluation Frameworks;
- inadequate description of methodological details and neglecting to append or reference the data collection instruments;
- inadequate utilization of performance monitoring data and the views of independent key informants with no stake in the program;
- inadequate assessment of incremental program impacts and, related to this, insufficient use of comparison groups and baseline measures in evaluation designs; and
- superficial coverage of cost-effectiveness issues.

Highlights of the findings for each major issue/requirement assessed in the review are as follows:

- **Executive Summary:** Although most reports (86 per cent) included an Executive Summary, the summaries are in need of some improvement. One-quarter of those reviewed were rated as inadequate[6] as a coherent, stand-alone document and approximately one-third lacked any presentation of the evaluation issues – though this latter deficiency is less common in reports submitted after April 2002 (22 per cent) than before (56 per cent).
- **Introduction and Context:** Most of the evaluation reports reviewed provided a good presentation of the program/initiative being evaluated, including its resources, beneficiaries and stakeholders. In addition, about six in ten reports discussed the underlying assumptions of the program (e.g., funding, partnerships), external factors such as environmental influences, and the timing and significance of the evaluation. Most reports also included a clear statement of the objectives of the evaluation. On the other hand, most reports lacked a presentation or reference to a logic model and a discussion of the major cause and effect relationships upon which the program was based (less than one-quarter of the evaluations included these elements). Most of the reports only listed (two-thirds) and very few discussed (about one-quarter) the evaluation issues. Moreover, half of the reports did not reference any document, such as an RMAF or Evaluation Framework, as context for the development of the evaluation issues.
- **Methodology:** The majority of evaluations (72 per cent) employed an appropriate research design, in light of the study's objectives. Only five per cent were found not to have an appropriate design (e.g., because they consulted very few respondents or included a limited range of perspectives), though we were unable to make an assessment on this criterion for

6 Most of the criteria assessed in this review were rated on a five-point scale ranging from 1 ("poor") to 5 ("excellent"), with the mid-point 3 indicating "adequate". In the presentation of findings in this chapter, the scale ratings are collapsed into the three following categories: 1-2 ("inadequate"), 3 ("adequate") and 4-5 ("more than adequate").

almost one-quarter (23 per cent) of the reports due to a lack of details. Among those reports assessed, the quality of the methodological design was rated as adequate or better for 87 per cent of evaluations. Virtually all of the evaluations (97 per cent) included multiple lines of evidence. There were also weaknesses, however. Many reports lacked a full description of the key methodological details. While just over half of reports described the methodology, four in ten only listed a few details. Only one-quarter of reports referenced a technical document with more methodological details. Consequently, 46 per cent of the reports were rated as inadequate in their methodological description. Moreover, half of the reports included no data collection instruments or a reference to where the instruments could be found. Only a minority of evaluations incorporated data from a performance measurement system (24 per cent) or from interviews with independent key informants with no stake in the program (26 per cent). This latter feature is, however, more common in evaluations completed after April 2002 than those done earlier (31 versus 16 per cent). Only a minority of the evaluation designs included a comparison group (13 per cent), baseline measures (14 per cent) or a comparison to norms, literature or some other benchmark (22 per cent) – features that can enhance the rigour of the methodology. Finally, only about four in ten evaluation reports included a statement of the limitations or constraints of the evaluation.

- **Findings – Relevance:** Over half of the evaluations (just under 60 per cent) provided a presentation of findings related to the continuing need for and relevance of the program. Of these evaluations, the majority (85 per cent) were rated as adequate or more than adequate on these criteria. Only about one-third of evaluations presented findings on whether the program duplicates or works at cross purposes with other programs/initiatives, and among those that did, this presentation was rated as inadequate for 18 per cent.

- **Findings – Success:** The majority of evaluations (87 per cent) reported findings demonstrating whether or not the program/initiative in question was producing results that supported its continuation or renewal. Although roughly one-quarter of these reports (26 per cent) were rated as inadequate on this criterion, the proportion with a less-than-adequate presentation of these results has decreased (19 per cent post-April 2002 versus 39 per cent pre-April 2002). Only one-quarter of the evaluations discussed unintended outcomes (25 per cent) or addressed incremental impacts (26 per cent). Neither of these issues was addressed in roughly two-thirds of the evaluations.

- **Findings – Cost-Effectiveness:** Only 26 per cent of evaluations provided findings on alternative, potentially more cost-effective approaches, though coverage of this issue has increased in more recent reports (31 per cent post-April 2002 versus 16 per cent pre-April 2002). In addition, roughly one-third of the evaluations (34 per cent) provided a qualitative and/or quantitative assessment of the cost-effectiveness of the program/initiative under review, though 28 per cent of these evaluations were rated as inadequate on this criterion.

- **Findings – Delivery/Implementation:** With respect to delivery/implementation issues, most evaluations presented findings related to the appropriateness of the program's delivery model and/or management practices (81 per cent) and the need to improve the program structure or delivery arrangements (76 per cent). The evaluations were rated highly on the former criterion (89 per cent adequate or more than adequate).

- **Findings – Appropriateness of Analysis:** It was difficult to assess the appropriateness of the analysis (i.e., the degree to which the analysis was supported by the data as determined by significance tests, response rates, etc.) for 50 per cent of the evaluations due to a lack of details in the reports. Among the reports that were assessed on this criterion, almost one-third (32 per cent) were rated as inadequate. This latter proportion has, however, decreased in recent years (26 per cent post-April 2002 compared to 41 per cent pre-April 2002).

- **Conclusions:** Three-quarters of evaluations were rated as adequate or better, and one-quarter (24 per cent) as inadequate, in their provision of objective, evidence-based conclusions related to relevance, success and/or cost-effectiveness. Among evaluations that addressed implementation/delivery and/or management practices, a higher proportion (85 per cent) were rated as adequate or better in providing objective, evidence-based conclusions on these issues. Moreover, the quality of evaluations is improving on this criterion: 40 per cent of the evaluations completed after April 2002 were rated as more than adequate in this regard compared to only 20 per cent of reports completed before this time. In addition, in their conclusions, half of the evaluations (49 per cent) presented other lessons learned about the program. Among these reports, 95 per cent were rated as adequate or more than adequate on this point.

- **Recommendations:** The vast majority of evaluations included formal recommendations (77 per cent) or suggestions for further action (13 per cent). In almost all cases, the recommendations addressed significant evaluation findings and flowed logically from the findings and conclusions (94 per cent in each case). On the other hand, among the reports with recommendations, only 26 per cent identified alternative scenarios and just 35 per cent took practical constraints (e.g., regulations, budgets) into account. Over one-third of these reports (35 per cent) were rated as inadequate on this criterion.

- **Management Response and Action Plan:** Less than half of the evaluation reports included a management response (48 per cent) or action plan (33 per cent).

- **General/Other Aspects of Report:** Most evaluation reports were rated as adequate or more than adequate in terms of being clearly written (86 per cent) and well-organized (81 per cent). Regarding weaknesses, a substantial proportion of the reports were rated as inadequate with respect to the fair presentation of data, including numbers and sources (33 per cent), the appropriate presentation of technical information (30 per cent), and the effective use of tables and charts (25 per cent).

- **Overall Assessment:** The majority of evaluation reports received an overall subjective rating of adequate (45 per cent) or more than adequate (32 per cent), though almost one-quarter of the evaluations (23 per cent) were rated as inadequate.

3.2 Detailed Findings

A) Executive Summary

The majority of reports reviewed (86 per cent) included an Executive Summary. Departments were more likely to include an Executive Summary in their evaluation reports than agencies (90 versus 71 per cent). Also, large and medium-sized organizations (83 and 92 per cent, respectively) were more likely to include a summary than small organizations (78 per cent).

With respect to being clearly and concisely written as well as coherent as a stand-alone document, most of the Executive Summaries were rated as adequate or more than adequate (43 and 31 per cent, respectively), whereas one-quarter were rated as inadequate.

Other key features of the Executive Summaries are as follows:

- Key evaluation issues were presented completely (38 per cent) or partially (30 per cent) in most Executive Summaries, though not at all in 32 per cent of the report's summaries. Executive Summaries that lacked a presentation of the evaluation issues were more common in reports submitted prior to April 2002 than later (56 versus 22 per cent) and in reports from small organizations (57 per cent) than those from large and medium-sized organizations (31 and 26 per cent respectively).

- Key evaluation findings were summarized in almost all Executive Summaries, either completely (50 per cent) or partially (43 per cent).

- Key conclusions were also summarized in most Executive Summaries, either completely (60 per cent) or partially (26 per cent).

- Evaluation recommendations were presented completely (69 per cent) or partially (nine per cent) in a majority of report Executive Summaries.

B) Introduction and Context

Description

The vast majority of federal program evaluations – 98 per cent – provided a clear and concise description of the program, policy or initiative being evaluated (see Table 1). The ratings of the quality of the program description were also strong: 35 per cent of evaluations were rated as

adequate on this criterion and another 49 per cent of evaluations provided a more than adequate discussion.

Most reports described all (64 per cent) or some (29 per cent) of the intended beneficiaries and stakeholders involved in the program, policy or initiative. The majority of reports were rated adequate (61 per cent) or more than adequate (25 per cent) on this criterion. Evaluation reports were somewhat more likely to have identified the beneficiaries of the program (77 per cent) than its stakeholders (68 per cent).

The majority of federal evaluation reports (71 per cent) included a discussion of resource allocation in the program description. Among these reports, this discussion was rated adequate (37 per cent) or more than adequate (40 per cent).

About six in ten (59 per cent) federal evaluation reports provided a description of the underlying assumptions of the program under study (e.g., funding, partnerships) or external factors (e.g., environmental influences). Of those reports (n=68) that identified these factors, 78 per cent described underlying assumptions of the program, while 66 per cent identified external factors.

The key weakness in the program description component was the lack of reference to a program logic model: fewer than one in four federal evaluation reports presented a logic model (19 per cent in the report itself and another four per cent in a referenced document). Related to this, just 22 per cent of federal evaluation reports included a description of the major cause and effect relationships upon which the program or policy was based (e.g., as presented in the logic model). Of those reports that included a discussion of the major cause and effect relationships (n=29), the discussion was rated adequate or more than adequate for most (41 and 31 per cent, respectively), but inadequate for 28 per cent of the reports.

TABLE 1: Program Description – Criteria and Ratings				
Criteria	**Met Criteria (%)**	**Ratings**		
		Inadequate (%)	**Adequate (%)**	**More Than Adequate (%)**
Describes program, policy, initiative	98	16	35	49
Describes beneficiaries/stakeholders	93*	14	61	25
Discusses resource allocation	71	23	37	40
Describes underlying assumptions/external factors	59	10	59	30
Presents logic model	23**	n/a	n/a	n/a
Describes major cause and effect relationships	22	28	41	31

Source: Review of Federal Program Evaluations (n=115). Only reports that met criteria were subject to ratings (n=29 to 113).“n/a” indicates that no rating was made on a criterion.*All or some beneficiaries.**Presented in report or reference to other document.

Evaluation Context

The vast majority of federal evaluation reports (91 per cent) included a statement regarding the objectives of the evaluation (Table 2). The quality rating was high for this criterion: 52 per cent of reports received a rating of adequate and another 32 per cent were rated more than adequate in this area.

About six in ten reports (58 per cent) provided an indication of the timing of the evaluation (i.e., the period over which the study took place) and a similar proportion (56 per cent) described the significance of the evaluation. A discussion of the evaluation's significance was more common in reports from departments than agencies (59 versus 42 per cent) and in those from large organizations (65 per cent) than medium-sized or small organizations (53 and 39 per cent, respectively). The rated quality of this criterion was positive: 30 per cent rated more than adequate, 59 per cent adequate and 11 per cent inadequate.

In terms of evaluation issues and questions, the typical practice in federal evaluation reports (two-thirds) is to merely list the questions (as opposed to discussing them, which was observed in just 24 per cent of the reports). The rating of this criterion was comparatively low in relation to other scores owing to this practice. On this criterion, 45 per cent of reports received an adequate rating and 20 per cent were more than adequate, whereas 35 per cent were rated as inadequate.

A small minority of federal evaluation reports (eight per cent) identified the evaluation issues within the context of a Results-based Management and Accountability Framework (RMAF). There was virtually no difference on this item based on when the evaluation was completed (pre- or post-April 2002). However, 42 per cent of reports discussed the issues and questions within

the context of another document (typically an Evaluation Framework). Half of the reports did not reference any context for the development of the evaluation issues and questions.

TABLE 2: Evaluation Context — Criteria and Ratings

Criteria	Met Criteria (%)	Ratings		
		Inadequate (%)	Adequate (%)	More Than Adequate (%)
Describes objectives of the evaluation	91	16	52	32
Describes timing of evaluation	58	n/a	n/a	n/a
Describes significance of evaluation	56	11	59	30
Describes issues and questions	89*	35	45	20
Describes timing of evaluation	58	n/a	n/a	n/a

Source: Review of Federal Program Evaluations (n=115). Only reports that met criteria were subject to ratings (n=64 to 106). "n/a" indicates that no rating was made on a criterion.* Describes or lists issues.

In terms of issue coverage7 (Table 3), the vast majority of federal evaluation reports covered success issues (94 per cent), followed by relevance (74 per cent) and implementation/delivery issues (72 per cent). Reports are far less likely to have addressed management practices (47 per cent) or cost-effectiveness (44 per cent).

The coverage of relevance issues was more common in evaluations from small and medium-sized organizations (89 and 80 per cent, respectively) than in those from large organizations (61 per cent). Cost-effectiveness issues were more likely to be addressed in evaluations completed after April 2002 than before (51 versus 27 per cent). Addressing issues related to management practices was more common in evaluations from departments than those from agencies (52 versus 29 per cent) and in reports from large and medium-sized organizations (50 and 51 per cent, respectively) than those from small organizations (28 per cent).

TABLE 3: Coverage of Evaluation Issues

Issue	Covers (%)
Relevance	74
Success	94
Cost-Effectiveness	44
Implementation/delivery	72
Management practices	47

Source: Review of Federal Program Evaluations (n=115)

7 Aside from the core TB evaluation issues of a program's continued relevance, results/success and cost-effectiveness, some evaluation reports covered issues related to the program's implementation/delivery (e.g., the degree to which the program's intended outputs were being produced and delivered to the intended beneficiaries) and management practices (e.g., the suitability of the program's governance structure, the clarity of management roles, responsibilities and communications).

C) Methodology

Description of Methodology/Design

Discussions of the evaluation methodology in federal evaluation reports were of varying quality – 56 per cent provided a full description of the methodologies and design applied to the evaluation (Table 4). Four in ten listed a few details only.

In the discussion of methodology, reports were most likely to identify sample size (e.g., of key informant interviews, surveys) (68 per cent). In terms of other elements, 45 per cent indicated the sampling method, 30 per cent linked methods to issues and 26 per cent provided data collection instruments. One-quarter of reports (27 per cent) referenced a technical document providing more details on the methodology. Three in ten reports contained none of the above elements in their methodological discussion (i.e., sample size, sample method, instruments, linking methods to issues, reference to technical documents).

The lack of methodological detail translated into a weak rating of the quality of reports in terms of this criterion: 46 per cent of reports were rated as inadequate on this item, 32 per cent of reports received an adequate rating and 21 per cent of reports were considered more than adequate.

Half of federal evaluation reports (49 per cent) did not include data collection instruments in the report, nor was there a reference to a technical document where the instruments could be located. This deficiency was more common in evaluations from medium-sized organizations (61 per cent) than in those from large or small organizations (37 and 44 per cent, respectively). One-quarter of reports (23 per cent) presented all research instruments with the report and another 10 per cent provided some of the instruments. Eighteen per cent of reports referenced a technical document where the instruments could be found.

On the whole, the majority of evaluations (72 per cent) were found to employ a design appropriate for the intended objectives of the study (based on such considerations as cost-effectiveness, feasibility, validity). Five per cent of evaluations did not meet this criterion and in 23 per cent of cases, the reviewer was unable to make an assessment (due to inadequate description of the design). Those that were considered to be inadequate tended to only have a limited range of perspectives represented (e.g., no client input, interviews with federal government representatives only) or to have consulted only small numbers of individuals/organizations.

The rating of the quality of the methodological design was favourable: of the evaluation reports that were rated, 45 per cent were given a rating of adequate and 42 per cent were rated as more than adequate in this area. Only 14 per cent of these evaluations were considered to be inadequate in terms of design.

TABLE 4: Methodology — Criteria and Ratings

Criteria	Met Criteria (%)	Ratings		
		Inadequate (%)	Adequate (%)	More Than Adequate (%)
Describes methodologies and design applied	56	46	32	21
Elements of Description		n/a	n/a	n/a
Sample Size	68			
Sample Method	45			
Links Methods to Issues	30			
Reference to Technical Documents	27			
Instruments	26			
Appropriate design	72	13	45	42

Source: Review of Federal Program Evaluations (n=115). Only reports that met criteria were subject to ratings (n=64 to 106). "n/a" indicates that no rating was made on a criterion.* Describes or lists issues.

Multiple Lines of Evidence

Among the strengths of federal program evaluations, virtually all studies (97 per cent) contained multiple lines of evidence to support findings (Table 5). Almost two-thirds of reports were rated as having an appropriate balance between qualitative and quantitative methodologies, while 14 per cent had an inappropriate balance (about two-thirds of these were considered to have been too reliant on qualitative methods) and in 23 per cent of cases the reviewer was unable to make an assessment.

The most frequently used methodologies were: key informant interviews (94 per cent), document reviews (78 per cent), sample surveys (38 per cent), file reviews (38 per cent), literature reviews (36 per cent), case studies (35 per cent), and focus groups (24 per cent).

Incorporation of data from an ongoing performance measurement system was infrequent: 24 per cent of reports indicated these data as a source of evidence for the evaluation.

A majority of reports were also rated to be of adequate (50 per cent) or more than adequate (28 per cent) quality in terms of inclusion of a variety of stakeholder perspectives. Federal program evaluations most often canvassed the perspective of program management and delivery personnel (83 per cent); clients/beneficiaries (58 per cent); partners (39 per cent); funding recipients (36 per cent); and third-party deliverers (24 per cent). In addition, experts were consulted in 20 per cent of the evaluations, and this was more common after April 2002 than before (24 versus 11 per cent).

In only 26 per cent of cases, however, was qualitative evidence drawn from key informants who did not have a stake in the program. This desirable methodological feature was more common in evaluations completed after April 2002 than earlier (31 versus 16 per cent), and in those from small and medium-sized organizations (39 and 33 per cent, respectively) than from large organizations (13 per cent).

TABLE 5: Multiple Lines of Evidence – Criteria and Ratings

Criteria	Met Criteria (%)	Ratings		
		Inadequate (%)	Adequate (%)	More Than Adequate (%)
Includes multiple lines of evidence	97	n/a	n/a	n/a
Use of ongoing performance monitoring data	24	n/a	n/a	n/a
Appropriate balance of qualitative and quantitative	64	n/a	n/a	n/a
Includes all stakeholder perspectives*	n/a	23	50	28
Non-stakeholder perspective included	26	n/a	n/a	n/a

Source: Review of Federal Program Evaluations (n=115)"n/a" indicates that no rating was made on a criterion *Only reports for which this criterion could be assessed were subject to this rating (n=97).

Limitations

Four in ten (39 per cent) federal program evaluation reports included a discussion of the limitations of the methodologies and data sources used (e.g., bias, data reliability). A similar proportion (44 per cent) indicated the constraints of the evaluation, with data availability and time (34 and 19 per cent, respectively) being the most often noted constraints.

Rigour

With respect to rigour, few federal program evaluations employed the traditional experimental or quasi-experimental design. While 44 per cent of evaluations included a representative survey of participants, only 13 per cent included a comparison group and 14 per cent compared evaluation data to a baseline measure. A somewhat larger proportion, 22 per cent, included comparative data from the literature or some other benchmark, however.

There is a trend for evaluations from medium-sized organizations to be somewhat less rigorous than those from large or small organizations. For example, a representative survey of participants and a comparison group were less common in evaluations in medium-sized organizations (31 and six per cent, respectively) than in those from small organizations (67 and 22 per cent) or large organizations (50 and 17 per cent).

D) Key Findings

Relevance

Just over half of the evaluation reports (57 per cent) presented evidence to demonstrate the actual need for the program in question as well as the program's responsiveness to this need (Table 6). The presentation of these findings was rated as being adequate or better for 85 per cent of the reports reviewed. Provision of evidence on these two issues was less common in reports from large organizations (46 and 48 per cent, respectively) than in those from medium-sized (61 and 59 per cent) or small organizations (78 per cent for both issues). Moreover, the quality of the evidence on the second issue (responsiveness to need) was rated differently according to size of organization. Reports from small and large organizations were more likely to be rated as more than adequate in this respect (47 and 41 per cent, respectively) than reports from medium-sized organizations (19 per cent). Note also that these issues were simply not addressed in roughly one-third of the evaluations.

TABLE 6: Relevance Findings — Criteria and Ratings

Criteria	Met Criteria (%)	Ratings		
		Inadequate (%)	Adequate (%)	More Than Adequate (%)
Evidence to demonstrate actual need	57	15	45	40
Evidence to demonstrate responsiveness to need	57	13	54	32
Evidence to demonstrate continued relevance to government priorities	58	12	47	41
Evidence to demonstrate does not duplicate	34	18	54	28

Source: Review of Federal Program Evaluations (n=115). Only reports that met criteria were subject to ratings (n=39 to 68).

Similarly, 58 per cent of the reports included evidence on the program's continuing relevance to government priorities and the presentation of these findings was rated as adequate (47 per cent) or more than adequate (41 per cent) for most reports. Again, however, the provision of evidence on this relevance issue was less common in reports from large organizations (48 per cent) than in those from medium-sized or small organizations (roughly two-thirds in each case). Fewer reports from large organizations were rated as more than adequate in this regard (30 per cent) than from small or medium-sized organizations (50 and 46 per cent, respectively). In addition, fewer reports submitted prior to April 2002 received a rating of more than adequate than those submitted after this time (32 versus 46 per cent). This issue was not addressed at all in 35 per cent of the evaluations.

Regarding the issue of whether the program duplicates or works at cross purposes with other programs/initiatives, only 34 per cent of the evaluations provided evidence and fully 54 per cent did not even address this issue. For the evaluations that did provide some evidence, the ratings were slightly lower than for the other relevance issues: 82 per cent of the reports were rated as adequate or better but 18 per cent were rated as inadequate on this point.

Success

The vast majority of evaluations (87 per cent) presented findings demonstrating whether or not the program, policy or initiative was producing results that supported its continuation or renewal (Table 7). Only four per cent of the evaluations did not present these success findings, and success issues were not addressed for the remaining nine per cent. The proportion that presented success findings was somewhat higher for small organizations (100 per cent) compared to medium-sized and large organizations (84 and 85 per cent, respectively).

TABLE 7: Success Findings — Criteria and Ratings

Criteria	Met Criteria (%)	Ratings		
		Inadequate (%)	Adequate (%)	More Than Adequate (%)
Describes program results/attribution of program to success	87	26	37	37
Identifies other programs, policies initiatives having relationships, shared results	37	n/a	n/a	n/a
Takes these into account in attribution	19	n/a	n/a	n/a
Discusses other factors contributing to results	61	14	50	36
Discusses unintended outcomes	25	14	60	21
Incrementality addressed	26	26	48	27

Source: Review of Federal Program Evaluations (n=115). Only reports that met criteria were subject to ratings (n=29 to 100) "n/a" indicates that no rating was made on a criterion.

About one-third (37 per cent) of the evaluations were judged to have described results more than adequately, a similar proportion (37 per cent) adequately, and 26 per cent, inadequately. The proportion indicating the presentation of findings was inadequate was considerably lower for large organizations (18 per cent) compared to small and medium-sized organizations (28 and 33 per cent, respectively); and for those produced after April 2002 than those produced before (19 versus 39 per cent).

A little over one-third of the evaluations (37 per cent) identified other programs, policies or initiatives that may have had similarities, relationships, shared results, and/or anticipated inter-program effects. About one-half (51 per cent) did not. The proportion that did not identify other programs was considerably higher for agencies (62 per cent) compared to departments (49 per cent).

About one-fifth of the evaluations (19 per cent) took other programs or initiatives into account in measuring success (attribution). Three in five (58 per cent) did not. The proportion taking other programs into account increases with the size of the organization, from six per cent for small organizations, 18 per cent for medium-sized ones, and 24 per cent for large organizations.

Three in five evaluations (61 per cent) discussed other factors that contribute to the results, while about one-third (31 per cent) did not. Smaller organizations (72 per cent) were more likely to consider other contributing factors than organizations in other size categories (59 per cent, for medium-sized and large organizations). In addition, agencies were considerably more likely to consider other factors than departments (75 versus 57 per cent). Similar proportions identified internal factors and external factors.

About one-third (36 per cent) of the evaluations were judged to have more than adequately considered other factors and 50 per cent to have done so adequately. Only 14 per cent were seen as considering contributory factors less than adequately. The proportion rated as more than adequate was considerably higher for medium-sized organizations (45 per cent) compared to small and large organizations (31 and 29 per cent).

One-quarter of evaluations (25 per cent) considered unintended outcomes and about two-thirds (63 per cent) did not. No significant differences emerged across the characteristics being considered. Of the evaluations that measured unintended outcomes, about half considered positive outcomes and about half considered negative outcomes.

About two-thirds of the evaluations (66 per cent) were seen as adequately discussing unintended outcomes, and one-fifth (21 per cent), more than adequately. There were too few observations to consider differences in results according to the size and type of organization and the timing of the evaluation.

One-quarter of evaluations (26 per cent) considered incrementally whereas almost two-thirds (64 per cent) did not. The measurement of incrementality was significantly higher for agencies than departments (38 versus 23 per cent), and for evaluations conducted after April 2002 than before (30 versus 17 per cent). Of the evaluations that did assess incrementality, 72 per cent looked at the issue subjectively, and 28 per cent, objectively. Incrementability was regarded as being adequately addressed in 53 per cent of the reports and more than adequately addressed in 27 per cent of them. There were too few observations to consider differences in results by the size and type of organization or by the timing of the evaluation.

Cost-Effectiveness

About one-quarter of the evaluations (26 per cent) discussed alternative approaches that could produce more cost-effective ways of achieving results. Sixteen per cent did not and for 58 per cent of the evaluations, cost-effectiveness was not addressed. The proportion of evaluations that did address alternative approaches declines steeply with the size of the organization, from 50 per cent for small organizations to 13 per cent for larger organizations. Also, this proportion is much larger in post-April 2002 evaluations than pre-April 2002 ones (31 versus 16 per cent), and somewhat larger in agencies compared to departments (38 versus 23 per cent).

Of the evaluations that considered alternative cost-effective approaches, 42 per cent were seen as assessing this adequately, and 29 per cent more than adequately. Again, there were too few observations to consider differences in results by size and type of organization or by the timing of the evaluation.

Of the evaluations that considered cost-effectiveness, about twice as many considered it qualitatively as quantitatively. This ratio did not vary much across the characteristics in question, apart from the ratio being somewhat lower in larger organizations. About one-half (49 per cent) of the qualitative or quantitative assessments of cost-effectiveness in the evaluations were considered to have been adequately carried out and one-quarter (23 per cent), more than adequately. Twenty-eight per cent of these evaluations were rated as inadequate, however. There two few observations to observe of how well cost-effectiveness was addressed across characteristics of organizations.

Delivery/Implementation

The majority of evaluations (81 per cent) presented findings related to appropriateness of the delivery model and/or management practices for contributing to the program's objectives. Specifically, almost two-thirds of the evaluations (64 per cent) assessed the delivery model and 50 per cent examined the management practices. An assessment of the latter issue was more common in reports from medium-sized and large organizations (55 and 52 per cent, respectively) than in those from small organizations (33 per cent). The presentation of these delivery/implementation findings was rated highly: 50 per cent of evaluations were regarded as adequate and 39 per cent as more than adequate. Ratings of greater than adequate were much higher in large organizations compared to small (43 versus 29 per cent).

In addition, most evaluations (76 per cent) presented evidence pertaining to the need to improve program structures or delivery arrangements. For 14 per cent of the evaluations reviewed, delivery/implementation issues were not addressed.

Other Aspects of Findings and Analysis

In most of the evaluations reviewed, the evaluation issues/questions were adequately (47 per cent) or more than adequately addressed (31 per cent), though 23 per cent of the evaluations were rated as inadequate on this criterion (see Table 8). In addition, with regard to the presentation of evidence-based findings that flow logically from the data and analyses, the majority of evaluations were rated as adequate or better (46 and 33 per cent, respectively) though about one-fifth (21 per cent) were seen as inadequate. Reports from small organizations were more likely to receive a rating of more than adequate on this point (44 per cent) than those from large or medium-sized organizations (36 and 26 per cent, respectively). Moreover, evaluations completed after April 2002 were somewhat more likely to be rated as more than adequate on this criterion than those done before this time (37 versus 24 per cent).

Regarding the appropriateness of the analysis (i.e., the degree to which the analysis is supported by the data as determined by significance tests, response rates, etc.), the ratings were fairly low. First, we were unable to make this assessment for 50 per cent of the evaluations – a finding that suggests that key details related to the analysis are not being included in evaluation reports. Second, among the reports that were assessed, about two-thirds were rated as adequate or better (47 and 21 per cent, respectively) but fully 32 per cent were regarded as inadequate on this key criterion. Reasons for considering analysis to have been inappropriate included: not attributing findings to specific distinct groups that had been consulted; not indicating the magnitude of a finding (e.g., the general proportion of stakeholders who may have held a certain view); relying too heavily on qualitative and anecdotal analysis; and presenting data with very small sample sizes without appropriate caveats. On a more encouraging note, fewer of the evaluations completed after April 2002 received a rating of inadequate than those done before this time (26 versus 41 per cent), suggesting some improvement. Ratings of inadequate were more frequent for agencies (55 per cent) compared to departments (26 per cent).

Table 8: Others Aspects of Findings and Analysis — Ratings

Criteria	Inadequate (%)	Adequate (%)	More Than Adequate (%)
Evaluation issues and questions are adequately addressed	23	47	31
Findings are based on the evidence and flow logically from the interpretation of data and analysis	21	46	33
Analysis is appropriate	32	47	21

Source: Review of Federal Program Evaluations (n=57 to 115)

E) Key Conclusions

The majority of evaluations presented conclusions on the relevance (57 per cent) and success (80 per cent) of the program/initiative in question, but only 29 per cent drew conclusions on cost-effectiveness. Note that fewer evaluations from large organizations presented conclusions on relevance or success (41 and 70 per cent, respectively) than those from small (67 and 89 per cent) or medium-sized organizations (67 and 86 per cent). Of the evaluations that drew conclusions on these three issues, most were rated as adequate (49 per cent) or more than adequate (27 per cent) with respect to the provision of objective, evidence-based conclusions though 24 per cent were seen as inadequate on this criterion (Table 9). Somewhat more evaluations from large organizations were rated as inadequate (31 per cent) than those from small or medium-sized organizations (about one-fifth in each case). Also, more evaluations completed after April 2002 received a rating of more than adequate on this criterion than those done earlier (30 versus 20 per cent), indicating some improvement.

TABLE 9: Conclusions — Criteria and Ratings

Criteria	Met Criteria (%)	Ratings		
		Inadequate (%)	Adequate (%)	More Than Adequate (%)
Provides objective, evidence-based conclusions on relevance, success and/or cost-effectiveness	n/a	24	49	27
Provides objective, evidence-based conclusions on implementation/delivery and/or management practices	n/a	15	52	33
Presents other lessons learned	54	5	54	41
Conclusions are based on explicit judgement criteria or benchmarks	21	n/a	n/a	n/a

Source: Review of Federal Program Evaluations (n=115). Only reports that met criteria were subject to ratings (n=56 to 96). "n/a" indicates that no rating was made on a criterion.

Almost two-thirds of the evaluations drew conclusions on implementation/delivery (64 per cent) but less than half addressed management practices in the conclusions (44 per cent). Conclusions on this latter issue were less common in evaluations from: small organizations (22 per cent) than large or medium-sized organizations (44 and 53 per cent, respectively); agencies than departments (33 versus 47 per cent); and after April 2002 than before this time (40 versus 54 per cent). The ratings for the provision of objective, evidence-based conclusions on these two issues were quite strong: the majority of evaluations were rated as adequate (52 per cent) or more than adequate (33 per cent). High ratings of more than adequate were more common for evaluations from large organizations (45 per cent) than small or medium-sized ones (roughly one-quarter in each case) and for those completed after April 2002 (40 per cent) than before this time (20 per cent).

About half of the evaluations (49 per cent) presented other lessons learned about the program. For these reports, the ratings were very high for this aspect. Just over half (54 per cent) were rated as adequate and fully 41 per cent were viewed as more than adequate. The highest ratings (i.e., more than adequate) were more common for evaluations completed after April 2002 than before (47 versus 25 per cent).

The evaluation conclusions were clearly based on explicit judgment criteria or benchmarks for only a minority (21 per cent) of the evaluations, though we were unable to make an assessment on this point for 34 per cent of the reports (e.g., due to a lack of information). A lack of such criteria/benchmarks was observed for 45 per cent of the evaluations overall, and this deficiency was more common for evaluations completed before April 2002 than later (57 versus 40 per cent).

F) Recommendations

Three-quarters of the evaluation reports reviewed contained formal recommendations (77 per cent). An additional 13 per cent contained suggestions for further action but these were not referred to as recommendations. Only 10 per cent of the reports did not contain any recommendations or suggestions. Formal recommendations were more like to appear in reports for small and medium-sized organizations (89 and 86 per cent, respectively) than in large organizations (63 per cent). Reports completed from April 2002 on were more likely to contain formal recommendations than those completed before (83 versus 65 per cent). Finally, reports completed for agencies were more likely to have formal recommendations than those done for departments (88 versus 75 per cent).

Of those reports with recommendations (n=99), 26 per cent identified alternative scenarios and 35 per cent took practical constraints such as regulations and budgets into account. While only 36 per cent were considered to be detailed, two-thirds were rated as operational (67 per cent) and just under two-thirds were evaluated as practical (61 per cent). Recommendations in reports from April 2002 and on were more likely to be operational and practical than earlier reports (72 versus 57 per cent and 65 versus 51 per cent, respectively). Recommendations in reports for agencies were more likely than those in reports for departments to be operational (79 versus 64 per cent).

Almost all of the reports with recommendations (94 per cent) addressed significant findings (i.e., key findings related to major, top priority evaluation issues), although nine per cent also addressed insignificant findings. As well, the vast majority of recommendations (94 per cent) were considered to flow logically from the findings and conclusions of the evaluation (Table 10).

One-quarter of reports with recommendations included a recommendation related to overall funding, and in all of these cases, the recommendation was to increase funding. Further, No reports presented evidence indicating that a program was not relevant or not needed. Any reports that presented evidence on relevance issues, presented evidence saying the evaluated program

was relevant and needed. However, one should note that these findings were sometimes accompanied by recommendations or suggestions that restructuring or other changes were needed, but always in context of program still being relevant/needed.

TABLE 10: Recommendations — Criteria and Ratings

Criteria	Met Criteria (%)	Ratings		
		Inadequate (%)	Adequate (%)	More Than Adequate (%)
Identifies alternative scenarios and takes into account any practical constraints	n/a	35	48	17
Recommendations are detailed and operational (and practical)	n/a	20	51	29
Recommendations address significant findings	94	13	57	30
Recommendations flow logically from findings and conclusions	94	15	53	32
Includes a recommendation related to overall funding	25	n/a	n/a	n/a

Source: Review of Federal Program Evaluations (n=115). Only reports that met criteria were subject to ratings (n=99 to 103). "n/a" indicates that no rating was made on a criterion.

G) Management Response and Action Plan

Just under half of the evaluation reports reviewed contained a management response (48 per cent). The remaining 52 per cent did not include this.

One-third of the evaluation reports reviewed contained an action plan in response to the evaluation (33 per cent). The remaining 67 per cent did not contain this element.

H) Clarity and Other Aspects of Report

In general, the evaluation reports were considered to have been clearly written, with 42 per cent considered to have been adequate and 44 per cent rated as more than adequate (Table 11). A full 17 per cent received a rating of excellent on this attribute. Twenty-two per cent of the reports contained a glossary of acronyms, to contribute to clarity. Reports submitted from April 2002 and later received higher ratings than those submitted before this date (53 versus 24 per cent were considered more than adequate).

With respect to the presentation of technical information, 55 per cent of the reports had sufficient but not excessive information in the body of the report and 38 per cent had relevant and supportive technical information in appendices (note that these two aspects are not mutually

exclusive). One-third of the reports (33 per cent), however, were considered to have been inadequate in terms of the appropriateness of the presentation of technical information.

Where there were technical appendices included with the report (n=72), the vast majority were considered to be of good quality (69 per cent adequate and 18 per cent more than adequate).

Forty-three per cent of the evaluation reports reviewed were between 25 and 40 pages, a length considered to be reasonable for the purposes to which these reports are put. Another 20 per cent were shorter than 25 pages and 37 per cent were longer.

Table 11: Clarity and Others Aspects of Report — Ratings

Criteria	Inadequate (%)	Adequate (%)	More Than Adequate (%)
Clearly written evaluation report	15	42	44
Appropriate presentation of technical information	30	51	18
Technical appendices are of high quality	13	69	18
Data are presented fairly	33	46	21
Effective use of tables and charts	25	52	23
Report is well- organized and easy to follow	19	49	32

Source: Review of Federal Program Evaluations (n=72 to 115)

Reports tended to do only moderately well in the context of the presentation of data. One-third were considered to have been inadequate with respect to the fair presentation of data (33 per cent), and 25 per cent were similarly rated as inadequate in terms of the effective use of tables and charts. On both of these attributes, just under one-quarter of reports were considered to be more than adequate. The largest proportion of reports, however, were considered to have been adequate both in terms of the fair presentation of data and the effective use of tables and charts (46 and 52 per cent respectively). Further, despite the above moderate ratings, 65 per cent of the reports provided numbers and 71 per cent documented the sources of the data.

Finally, in terms of whether the report was well-organized and easy to follow, almost one-third received a rating of more than adequate (33 per cent) and almost one-half were rated as adequate (49 per cent). Reports submitted from April 2002 and on were more likely to have been considered more than adequate on this attribute than those submitted before this date (39 versus 16 per cent).

I) Overall Assessment

At the end of each review, the reviewer gave the evaluation report a subjective rating of its overall quality. Most of the evaluation reports reviewed were considered to be adequate (45 per cent) or more than adequate (32 per cent), although only eight per cent were rated as "excellent". On the other hand, just under one-quarter (23 per cent) were judged as being, overall, inadequate.

There was no clear pattern to differences in the overall assessment as a function of organizational size (or example, reports from small organizations were both more likely to be rated as inadequate and as more than adequate than those from large organizations, whose reports were more likely to be judged as adequate than those from small organizations). Reports were more likely to be judged as inadequate, however, if submitted prior to April 2002 (32 per cent, compared to 18 per cent for April 2002 and beyond) and more like to be judged as more than adequate if submitted from April 2002 and beyond (37 per cent versus 22 per cent of those submitted prior to this date).

3.3 Strengths and Weaknesses of Federal Evaluations

A) Strengths

The key strengths of federal evaluations identified in this review are summarized below:

- Most of the evaluation reports reviewed provided a good presentation of the program/initiative being evaluated, including its resources, beneficiaries and stakeholders. About six in ten reports discussed the underlying assumptions of the program (e.g., funding, partnerships) and external factors such as environmental influences. Most reports also included a clear statement of the objectives of the evaluation.

- The majority of evaluations (72 per cent) employed an appropriate research design, in light of the study's objectives, though we were unable to make an assessment on this criterion for almost one-quarter of the reports due to a lack of details. Among those reports assessed, the quality of the methodological design was rated as adequate or better for 87 per cent of evaluations. Virtually all of the evaluations (97 per cent) included multiple lines of evidence.

- Over half of the evaluations (just under 60 per cent) provided a presentation of findings related to the continuing need for and relevance of the program in question. Of these evaluations, the majority (85 per cent) were rated as adequate or more than adequate on these criteria.

- The majority of evaluations (87 per cent) reported findings demonstrating whether or not the program/initiative in question was producing results that supported its continuation or renewal. Although roughly one-quarter of these reports (26 per cent) were rated as inadequate

on this criterion, the proportion with a less-than-adequate presentation of these results has decreased (19 per cent post-April 2002 versus 39 per cent pre-April 2002).

- With respect to delivery/implementation issues, most evaluations presented findings related to the appropriateness of the program's delivery model and/or management practices (81 per cent) and the need to improve the program structure or delivery arrangements (76 per cent). The evaluations were rated highly on the former criterion (89 per cent adequate or more than adequate).

- Among the evaluations addressing these issues, the majority (85 per cent) were rated as adequate or better in providing objective, evidence-based conclusions related to implementation/delivery and/or management practices. Moreover, the quality of evaluations is improving on this criterion: 40 per cent of the evaluations completed after April 2002 were rated as more than adequate in this regard compared to only 20 per cent of reports completed before this time.

- In their conclusions, half of the evaluations (49 per cent) presented other lessons learned about the program. Among these reports, 95 per cent were rated as adequate or more than adequate on this point.

- The vast majority of evaluations included formal recommendations (77 per cent) or suggestions for further action (13 per cent). In almost all cases, the recommendations addressed significant evaluation findings (i.e., key findings relating to the major evaluation issues) and flowed logically from the findings and conclusions (94 per cent in each case).

- Most evaluation reports were rated adequate or more than adequate in terms of being clearly written (86 per cent) and well-organized (81 per cent).

B) Weaknesses

The major weaknesses or areas in need of improvement in the federal evaluations included in this review are as follows:

- Executive summaries are in need of some improvement. One-quarter of those reviewed were rated as inadequate as a coherent, stand-alone document and approximately one-third lacked any presentation of the evaluation issues – though this latter deficiency is less common in reports submitted after April 2002 (22 per cent) than before (56 per cent).

- Most reports lacked a presentation or reference to a logic model and a discussion of the major cause and effect relationships upon which the program was based (less than one-quarter of the evaluations included these elements).

- Although about six in ten evaluation reports indicated the timing and significance of the evaluation, it would seem that a higher proportion of reports should include such basic details.

- Most of the reports only listed (two-thirds) and very few discussed (about one-quarter) the evaluation issues. Moreover, half of the reports did not reference any document, such as an RMAF or Evaluation Framework, as context for the development of the evaluation issues.

- Less than half of the evaluation reports (44 per cent) addressed cost-effectiveness issues, though coverage of these issues is more common in evaluations completed after April 2002 than before (51 versus 27 per cent).

- Many reports lacked a full description of the key methodological details. While just over half of reports described the methodology, four in ten only listed a few details. Only one-quarter of reports referenced a technical document with more methodological details. Consequently, 46 per cent of the reports were rated as inadequate in their methodological description. Moreover, half of the reports included no data collection instruments or a reference to where the instruments could be found.

- Only a minority of evaluations incorporated data from a performance measurement system (24 per cent) or from interviews with independent key informants with no stake in the program (26 per cent). This latter feature is, however, more common in evaluations completed after April 2002 than those done earlier (31 versus 16 per cent).

- Despite the fact that almost three-quarters of the evaluations were judged to have an appropriate research design for the study's objectives, only a minority of the evaluation designs included features to optimize the rigour of the research such as a comparison group (13 per cent), baseline measures (14 per cent) or a comparison to norms, literature or some other benchmark (22 per cent).

- Only about four in ten evaluation reports included a statement of the limitations or constraints of the evaluation.

- Only about one-third of evaluations presented findings on whether the program duplicates or works at cross purposes with other programs/initiatives.

- Only one-quarter of the evaluations discussed unintended outcomes (25 per cent) or addressed incremental impacts (26 per cent). Neither of these issues was addressed in roughly two-thirds of the evaluations.

- Only 26 per cent of evaluations provided findings on alternative, potentially more cost-effective approaches, though coverage of this issue has increased in more recent reports (31 per cent post-April 2002 versus 16 per cent pre-April 2002). In addition, roughly one-third of

the evaluations (34 per cent) provided a qualitative and/or quantitative assessment of the cost-effectiveness of the program/initiative under review, though 28 per cent of these evaluations were rated as inadequate on this criterion.

- It was difficult to assess the appropriateness of the analysis (i.e., the degree to which the analysis was supported by the data as determined by significance tests, response rates, etc.) for 50 per cent of the evaluations due to a lack of details in the reports. Among the reports that were assessed on this criterion, almost one-third (32 per cent) were rated as inadequate. This latter proportion has, however, decreased in recent years (26 per cent post-April 2002 compared to 41 per cent pre-April 2002).

- Almost one-quarter of evaluations (24 per cent) were rated as inadequate in their provision of objective, evidence-based conclusions related to relevance, success and/or cost-effectiveness.

- Among the reports with recommendations, only 26 per cent identified alternative scenarios and just 35 per cent took practical constraints (e.g., regulations, budgets) into account. Over one-third of these reports (35 per cent) were rated as inadequate on this criterion.

- Less than half of the evaluation reports included a management response (48 per cent) or action plan (33 per cent).

- Over one-third of the reports (37 per cent) were quite long – more than 40 pages in length.

- A substantial proportion of evaluation reports were rated as inadequate with respect to the fair presentation of data, including numbers and sources (33 per cent), the appropriate presentation of technical information (30 per cent), and the effective use of tables and charts (25 per cent).

3.4 Variations in Quality by Organizational Characteristics and Report Date

A) Size of Organization

A number of interesting differences were identified by the size of organization. However, there was no consistent pattern in the results by organization size. Organizations in any particular size categories were not judged consistently as having higher quality evaluations than organizations in other size categories. The major differences by size included the following:

- Large and medium-sized organizations (83 and 92 per cent, respectively) were more likely to include an Executive Summary than small organizations (78 per cent).

- Executive Summaries that lacked a presentation of the evaluation issues were more common in reports from small organizations (57 per cent) than those from large and medium-sized organizations (31 and 26 per cent, respectively).

- A discussion of the evaluation's significance was more common in reports from large organizations (65 per cent) than medium-sized or small organizations (53 and 39 per cent, respectively).

- The coverage of relevance issues was more common in evaluations from small and medium-sized organizations (89 and 80 per cent, respectively) than in those from large organizations (61 per cent). Addressing issues related to management practices was more common in reports from large and medium-sized organizations (50 and 51 per cent, respectively) than those from small organizations (28 per cent).

- Lack of data collection instruments in the report, and reference to a technical document where the instruments could be located, was more common in evaluations from medium-sized organizations (61 per cent) than in those from large or small organizations (37 and 44 per cent, respectively).

- Drawing qualitative evidence from key informants who did not have a stake in the program was more common in evaluations from small and medium-sized organizations (39 and 33 per cent, respectively) than from large organizations (13 per cent).

- A representative survey of participants and a comparison group were less common in evaluations in medium-sized organizations (31 and six per cent, respectively) than in those from small organizations (67 and 22 per cent) or large organizations (50 and 17 per cent).

- Reports from medium-sized organizations were much less likely to be rated as more than adequate in providing evidence on responsiveness to need (19 per cent) than reports from either small or large organizations (47 and 41 per cent, respectively).

- The provision of evidence on the continued relevance issue was less common in reports from large organizations (48 per cent) than in those from medium-sized or small organizations (roughly two-thirds in each case). Fewer reports from large organizations were rated as more than adequate in this regard (30 per cent) than from small or medium-sized organizations (50 and 46 per cent, respectively).

- The proportion that presented success findings was somewhat higher for small organizations (100 per cent) compared to medium-sized and large organizations (84 and 85 per cent, respectively).

- The proportion of evaluations rated as inadequate for the presentation of findings was considerably lower for large organizations (18 per cent) compared to small and medium-sized organizations (28 and 33 per cent, respectively).

- The proportion taking other programs into account in assessing impacts increases with the size of the organization, from six per cent for small organizations, 18 per cent for medium-sized ones, to 24 per cent for large organizations.

- Smaller organizations (72 per cent) were more likely to consider other contributing factors than organizations in other size categories (59 per cent, for medium-sized and large organizations).

- The proportion rated as more than adequate in considering contributing factors in measuring success was considerably higher for medium-sized organizations (45 per cent) compared to small and large organizations (31 and 29 per cent).

- The proportion of evaluations that assessed alternative approaches declines steeply with the size of the organization, from 50 per cent for small organizations to 13 per cent for larger organizations.

- An assessment of management practices was more common in reports from medium-sized and large organizations (55 and 52 per cent, respectively) than in those from small organizations (33 per cent). Ratings of more than adequate were much higher in large than small organizations (45 versus 29 per cent).

- For the presentation of evidence-based findings that flow logically from the data and analyses, more reports from small organizations received a rating of more than adequate (44 per cent) than those from large or medium-sized organizations (36 and 26 per cent, respectively).

- Conclusions on management practices were less common in evaluations from small organizations (22 per cent) than large or medium-sized organizations (44 and 53 per cent, respectively).

- High ratings of more than adequate for the provision of evidence-based conclusions on delivery and management practices issues were more common for evaluations from large organizations (45 per cent) than small or medium-sized ones (roughly one-quarter in each case).

- Formal recommendations were more like to appear in reports for small and medium-sized organizations (89 and 86 per cent, respectively) than in those for large organizations (63 per cent).

B) Pre- versus Post-April 2002

There were some key differences according to when the report was produced. In general, evaluations completed after April 2002 were rated more highly than those done earlier. The detailed results follow:

- Executive Summaries that lacked a presentation of the evaluation issues were more common in reports submitted prior to April 2002 than later (56 versus 22 per cent).
- Cost-effectiveness issues were more likely to be addressed in evaluations completed after April 2002 than before (51 versus 27 per cent).
- The presentation of qualitative evidence drawn from key informants who did not have a stake in the program was more common in evaluations completed after April 2002 than earlier (31 versus 16 per cent).
- Fewer reports submitted prior to April 2002 received a rating of more than adequate on the findings for continuing relevance than those submitted after this time (32 versus 46 per cent).
- The proportion for which the presentation of findings on success was inadequate was considerably lower in reports produced after April 2002 than those produced before (19 versus 39 per cent).
- The proportion of evaluations that addressed alternative approaches was much larger in post-April 2002 evaluations than pre-April 2002 ones (31 versus 16 per cent).
- Evaluations completed after April 2002 were somewhat more likely to be rated as more than adequate for the presentation of evidence-based findings that flow logically from the data and analyses than those done before this time (37 per cent versus 24 per cent).
- Fewer evaluations completed after April 2002 were rated as inadequate with respect to the appropriateness of the analysis than those done earlier (26 versus 41 per cent).
- More evaluations completed after April 2002 received a rating of more than adequate with regard to the provision of objective, evidence-based conclusions (on relevance, success and/or cost-effectiveness) than those done earlier (30 versus 20 per cent), indicating some improvement.
- Conclusions on management practices were less common in evaluation reports dated after April 2002 than before this time (40 versus 54 per cent). High ratings of more than adequate for the conclusions on delivery/implementation issues were more common for evaluations completed after April 2002 (40 per cent) than before this time (20 per cent).

- Reports completed from April 2002 on were more likely to contain formal recommendations than those completed before (83 versus 65 per cent).

- Reports submitted from April 2002 were more likely than those submitted before this date to be considered more than adequately clearly written (53 versus 24 per cent).

- Reports were more likely to be rated as inadequate overall if submitted prior to April 2002 (32 per cent, compared to 18 per cent for April 2002 and beyond) and more likely to be judged as more than adequate if submitted April 2002 or later (37 per cent versus 22 per cent of those submitted prior to this date).

C) Agency Versus Department

Some differences were observed between evaluations sponsored by agencies and those sponsored by departments, but there was no consistent pattern in the results. The differences by agency and department include the following:

- A discussion of the evaluation's significance was more common in reports from departments than agencies (59 versus 42 per cent).

- Addressing issues related to management practices was more common in evaluations from departments than those from agencies (52 versus 29 per cent).

- Drawing evidence from key informants without a stake in the program was more prevalent in post-April 2002 evaluations than earlier (31 versus 16 per cent).

- Evaluations from agencies were considerably more likely to consider other factors contributing to results than departments (75 versus 57 per cent).

- The measurement of incrementality was included in more evaluations for agencies than departments (38 versus 23 per cent).

- Evaluations from agencies were more likely to address alternative approaches than those from departments (38 versus 23 per cent).

- Conclusions drawn on implementation/delivery were less common in evaluations sponsored by agencies than departments (33 versus 47 per cent).

- Reports completed for agencies were more likely to have formal recommendations than those for departments (88 versus 75 per cent).

- Recommendations in reports for agencies were more likely than those for departments to be operational (79 versus 64 per cent).

4. Conclusions and Recommendations

4.1 Conclusions

On balance, most evaluations that were assessed in this review are of reasonable quality. The majority received an overall rating of adequate (45 per cent) or more than adequate (32 per cent). Still, a considerable proportion of the evaluations (23 per cent) were rated as inadequate and this finding warrants attention by the CEE. No clear, consistent patterns were observed when we compared the reports from organizations of different sizes or those from departments versus agencies. A noticeable improvement on a number of criteria was observed, however, when we compared evaluations completed prior to April 2002 with those done after this point in time. The latter, more recent evaluations show a significant improvement in quality, suggesting that TBS's April 2001 Evaluation Policy may have had a favourable impact.

As was detailed in the previous chapter, a number of strengths were identified in federal program evaluations. Key strengths include: a comprehensive description of the program/initiative under review including its resources, beneficiaries and stakeholders; a clear statement of the evaluation objectives; the use of multiple lines of evidence in the methodology; a strong presentation of findings, in particular, on relevance and delivery/implementation issues; inclusion of formal recommendations or suggested improvements, with the recommendations flowing logically from the findings and conclusions; and reports that are well-written and well-organized.

On the other hand, a number of weaknesses of evaluations and reports were also revealed by this review, including the following: neglecting to present or reference the program logic model; inadequate discussion of the evaluation issues and failing to reference source documents such as RMAFs or Evaluation Frameworks; inadequate description of methodological details and neglecting to append or reference the data collection instruments; inadequate utilization of performance monitoring data and the views of independent key informants with no stake in the program; inadequate assessment of incremental program impacts and insufficient use of comparison groups and baseline measures in evaluation designs; and superficial coverage of cost-effectiveness issues.

4.2 Recommendations

On the basis of the findings of this review, it is recommended that the CEE:

❶ Encourage evaluation divisions in federal departments and agencies to strengthen their evaluation reports by addressing the major weaknesses identified in this review:

Improving Evaluation Reports

- ensure that a report's Executive Summary includes all key points and serves as a stand-alone summary of the evaluation's objectives, issues, methodological approach, key findings, conclusions and (if applicable) recommendations;
- present the program logic model in the report or an appendix, or provide a reference where it can be found (e.g., RMAF, Evaluation Framework);
- list all evaluation issues in the report or an appendix, or provide a reference for the full list;
- provide all key details of the methodology (e.g., methods used, timing of data collection, number of respondents, types of analysis) and the data collection instruments, either in the report and its appendices or in a technical document that is referenced;
- state the limitations of and constraints on the evaluation;
- present the findings/data fairly by including key details on the data and analysis in the report or appendices, in particular, response rates, significance tests, numbers/quantitative results and sources of data;
- present objective, evidence-based conclusions, which are clearly and logically linked to the evaluation findings on which they are based;
- in the recommendations, consider alternate scenarios (if applicable) and practical constraints on suggested courses of action;
- try to keep the body of the evaluation report to 25-40 pages in length, and present essential supplementary information (e.g., detailed findings and technical analyses) in appendices;

Improving Evaluation Methodologies

- consult independent key informants (with no stake in the program) in more evaluations;
- incorporate an analysis of performance monitoring data in more evaluations;
- incorporate baseline measures and a comparison group in the research design for evaluations where incremental program impacts are an important issue; and
- include a quantitative assessment of cost-effectiveness issues in more final/summative evaluations.

❷ Refine Treasury Board guidelines/criteria for the expected features of (1) evaluation methodologies and (2) evaluation reports and disseminate them;

❸ Continue to implement a rigorous approach to monitoring the quality of evaluations, and use this as a basis for the development of individual report cards on the quality and overall health of the evaluation function by department and small agency; and,

❹ Identify measures, including an incentive structure and standards, to ensure that departments and agencies submit completed evaluations and reviews in a responsible, reasonable manner. Departments and agencies adherence to such standards should be made a public record.

Appendix A
Review Template

Review of the Quality of Evaluations

Review Template
(Final version: Draft 7)

Evaluation Report Description

Report ID Number			
Department	Small ❑	Medium ❑	Large ❑
Agency	Small ❑	Medium/Large ❑	
Size of Org. Evaluation Group			
Type of Report	❑ Review		
	❑ Formative Evaluation		
	❑ Summative Evaluation		
	❑ Special Study (e.g., research)		
	❑ Other: _____		
Date of Report			
Reviewer			

Review of the Quality of Evaluations
Review Template (April 26, 2004)

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
1.0 Executive Summary (Note: Assess Last)			☐ Yes ☐ No				
	1.1 Clearly and concisely written, coherent as a stand alone document				Poor 1 2 Adequate 3 4 Excellent 5		
	1.2 Presents key evaluation issues and answers these issues with relevant information through sound analysis	Key evaluation **issues** are summarized	☐ Yes – completely ☐ Yes – partially ☐ No				
		Key evaluation **findings** are summarized	☐ Yes – completely ☐ Yes – partially ☐ No				
		Key evaluation **conclusions** are summarized	☐ Yes – completely ☐ Yes – partially ☐ No				
		Evaluation **recommendations** are presented	☐ Yes – completely ☐ Yes – partially ☐ No ☐ N/A				

8 A rating of 3 indicates that the criterion is met, while a rating of 1or 2 indicates that it is not adequately met. A rating of 4 or 5 indicates excellent quality whereby the basic, minimum considerations for the criterion are exceeded or extremely well done.

9 Qualitative assessment to be completed only when a ✓ appears in the cell.

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
2.0 Introduction and Context							
2.1 Description	2.1.1 Clearly and concisely describes the program, policy or initiative being evaluated.		☐ Yes ☐ No		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
	2.1.2 Describes intended beneficiaries and stakeholders involved		☐ Yes – all ☐ Yes – some ☐ No	☐ beneficiaries ☐ stakeholders	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
	2.1.3 Describes the cause-and-effect linkages among inputs, activities, outputs, and outcomes, and external factors contributing to success or failure	Presents a logic model in report	☐ Yes ☐ No – but reference provided ☐ No – no reference				
		Major cause and effect relationships (e.g., as presented in the logic model) are described	☐ Yes ☐ No		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
		Underlying assumptions (e.g., funding, partnerships) and/or external factors (e.g., environmental influences) are described	☐ Yes ☐ No	☐ underlying assumptions ☐ external factors	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
	2.1.4 Discusses resource allocation to policy, program or initiative areas	Program resources are clearly described so that one understands how program funds have been allocated and spent	☐ Yes ☐ No		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
2.2 Evaluation Context	2.2.1 Identifies the role of the evaluation and its importance/ significance at the time it was conducted	Report describes the objectives of the evaluation	☐ Yes ☐ No		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
		Report describes the timing of the evaluation	☐ Yes ☐ No				
		Report describes the significance of the evaluation	☐ Yes ☐ No		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
	2.2.2 Describes the key evaluation issues and questions linked to the program, policy, or initiative	Describes evaluation issues and questions	☐ Yes – discusses issues ☐ Yes – only lists issues ☐ No	☐ presents issues in a technical appendix	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
		Identification of evaluation issues within context of RMAF or other key documents	☐ Yes – RMAF ☐ Yes – other documents ☐ No ☐ Unable to assess				
		Covers: › relevance › success › cost-effectiveness		☐ relevance ☐ success ☐ cost-effectiveness			
		Includes issues related to: › implementation/delivery › management practices		☐ implementation/ delivery ☐ management practices			

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
3.0 **Methodology**							
3.1 **Description of the Methodology/ Design**	3.1.1 Describes logical, valid, evidence-based methodologies that are linked to the evaluation issues explored OR there is a clear reference to a technical document for this information	Describes the methodologies and design applied to the evaluation	☐ Yes – describes ☐ Yes – only lists a few details ☐ No – no reference to technical documents ☐ No – reference to technical documents	☐ sample size ☐ sample method ☐ instruments ☐ links methods to issues ☐ reference to technical documents	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
		Instruments are presented	☐ Yes – all ☐ Yes – some ☐ No – no reference to technical documents ☐ No – reference to technical documents				
		The design is appropriate for the intended objectives of the study (e.g., cost-effective, feasible, logical, valid)	☐ Yes ☐ No ☐ Unable to assess		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9	✓	

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
3.2 **Multiple Lines of Evidence**	3.2.1 The evaluation contains multiple lines of evidence to support the validity of the findings	The evaluation relies on more than one line of evidence to support its findings › qualitative › quantitative › literature review › document review › file review › secondary data analysis › database review › analysis of performance data › case studies › cost-benefit analysis › other	☐ Yes ☐ No – but it should have ☐ No – but this is not necessary or appropriate for this evaluation	☐ qualitative ☐ focus group ☐ key informant interviews ☐ other _____ ☐ quantitative ☐ census ☐ sample survey ☐ other _____ ☐ literature review ☐ document review ☐ file review ☐ secondary data analysis ☐ database review ☐ analysis of performance data ☐ case studies ☐ cost-benefit analysis ☐ other _____			

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
		The evaluation uses data from an ongoing performance monitoring system	☐ Yes ☐ No – data available but not used ☐ No – data unavailable ☐ Not applicable ☐ Unable to assess				
	3.2.2 Is there an appropriate balance between qualitative and quantitative methodologies?		☐ Yes ☐ No ☐ N/A			✓	
	3.2.3 All stakeholder perspectives are included	› clients/beneficiaries › program management and delivery (federal government) › third-party deliverers › partners › experts › funding recipients › non-recipients › other _____	☐ Unable to assess	☐ clients/ beneficiaries ☐ program management and delivery (federal government) ☐ third-party deliverers ☐ partners ☐ experts ☐ funding recipients ☐ non-recipients ☐ other _____	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
		Qualitative evidence drawn from key informants who do not have a stake in the program	☐ Yes ☐ No ☐ Unable to assess				

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
3.4 Limitations	3.4.1 The limitations and trade-offs of the methodologies, data sources and data used in the evaluation are clearly articulated	Limitations are described: actual and potential biases, reliability of data are identified and explained in terms of their impact on stated findings	☐ Yes ☐ No ☐ No apparent limitations	☐ biases described ☐ data reliability explained	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
		The constraints of the evaluation are made clear	☐ Yes ☐ No ☐ No apparent constraints	☐ budget ☐ time ☐ data availability ☐ other _____	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
3.5 Rigour	3.5.1 A comparison "point" exists	Survey of representative group of participants	☐ Yes ☐ No				
		Comparison group	☐ Yes ☐ No				
		Comparison to baseline measures	☐ Yes ☐ No				
		Comparison to norms/literature/other benchmark	☐ Yes ☐ No				

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
4.0 Key Findings							
4.1 Relevance	4.1.1 Presents findings related to establishing continued relevance and contribution to results achievement by linking results to societal need and government priority areas	Evidence to demonstrate actual need	☐ Yes ☐ No ☐ Not addressed		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
		Evidence to demonstrate responsiveness to need	☐ Yes ☐ No ☐ Not addressed		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
		Evidence to demonstrate continued relevance to government priorities	☐ Yes ☐ No ☐ Not addressed		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
		Evidence to demonstrate that it does not duplicate or work at cross purposes with other programs, policies, or initiatives	☐ Yes ☐ No ☐ Not addressed		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
4.2 Success	4.2.1 Presents findings demonstrating whether or not the program, policy or initiative is producing results that support its continuation or renewal	Clearly describes what has happened as a result of the program and articulates attribution of program, policy or initiative to success	☐ Yes ☐ No ☐ N/A – success issues not addressed		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
	4.2.2 Identifies other programs, policies or initiatives that may have similarities, relationships, shared results, and/or anticipated inter-program effects	Identifies other programs, policies or initiatives	☐ Yes ☐ No ☐ N/A – success issues not addressed				
		Takes these into account in attribution	☐ Yes ☐ No ☐ N/A – success issues not addressed				
	4.2.3 Discusses other factors that contribute to the results (e.g., funding or partnering considerations, external factors)		☐ Yes ☐ No ☐ N/A – success issues not addressed	☐ factors internal to program ☐ external factors	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
	4.2.4 Discusses whether unintended outcomes were produced that have contributed to success or presented specific constraints		☐ Yes ☐ No ☐ N/A – success issues not addressed	☐ positive outcomes ☐ negative outcomes	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
	4.2.5 Incrementality is addressed		☐ Yes ☐ No ☐ N/A – success issues not addressed	☐ subjectively ☐ objectively	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
4.3 Cost-effectiveness	4.3.1 Identifies the extent to which the program, policy or initiative could have been delivered by more appropriate, cost-effective methods to achieve its objectives	Discusses alternative approaches that could produce more cost-effective ways of achieving results	☐ Yes ☐ No ☐ N/A – cost-effectiveness issues not addressed		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
		Presents: › qualitative assessment of cost-effectiveness › quantitative assessment of cost-effectiveness	☐ Yes ☐ No ☐ N/A – cost-effectiveness issues not addressed	☐ qualitative assessment ☐ quantitative assessment	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
4.4 Delivery/ Implementation	4.4.1 Presents findings related to identifying the efficacy and appropriateness of the scope of program structures and service delivery arrangements for the program, policy or initiative	Assesses delivery model and its appropriateness and contribution to meet objectives	☐ Yes	☐ delivery model	Poor 1		
		Provides evidence to identify whether there is a need for improved program structures or delivery arrangements	☐ Yes ☐ No ☐ N/A				
4.5 Evaluation Issues	4.5.1 The evaluation issues and questions are adequately addressed				Poor 1 2 Adequate 3 4 Excellent 5		
4.6 Evidence-based Findings	4.6.1 The findings are based on evidence drawn from the evaluation research	Demonstrates that the findings flow logically from the interpretation of the data and analyses			Poor 1 2 Adequate 3 4 Excellent 5		
4.7 Analysis	4.7.1 The analysis is appropriate	The data support the analysis (as determined by, for example, significance tests, response rates)	☐ Unable to assess		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9	✓	

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
5.0 Key Conclusions							
	5.1 Presents clear, impartial and accurate evidence-based conclusions	Conclusions objectively answer the evaluation issues and are supported by the findings		☐ relevance ☐ success ☐ cost-effectiveness	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
				☐ implementation/ delivery ☐ management practices	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
		Presents other lessons learned about the program from the evaluation	☐ Yes ☐ No ☐ Unable to assess		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
		Conclusions are based on explicit judgement criteria and benchmarks	☐ Yes ☐ No ☐ Unable to assess	☐ no criteria presented			

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
6.0 Recommendations			☐ Yes – formal recommendations ☐ Yes – suggestions that are not called "recommendations" ☐ No				
	6.1 Clearly states practical and realizable recommendations	Identifies alternative scenarios and takes into account any practical constraints (e.g., regulations, institutions, budget)		☐ alternative scenarios ☐ practical constraints	Poor 1 2 Adequate 3 4 Excellent 5		
		Recommendations are detailed and operational (and practical)		☐ detailed ☐ operational ☐ practical	Poor 1 2 Adequate 3 4 Excellent 5		
	6.2 Recommendations are supported by and flow logically from the findings and conclusions	Recommendations address significant findings	☐ Yes ☐ No	☐ also address insignificant findings	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
		Recommendations flow logically from findings and conclusions	☐ Yes ☐ No		Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
	6.3 Includes a recommendation related to overall funding		☐ Yes ☐ No	☐ increase funding ☐ decrease funding			
7.0 Management Response			☐ Yes ☐ No				
8.0 Action Plan			☐ Yes ☐ No				
9.0 General/Other							
9.1 Clarity	9.1.1 Report is written in plain language, and detailed technical information provided in technical appendices	Clearly written evaluation report		☐ glossary of acronyms	Poor 1 2 Adequate 3 4 Excellent 5		
		Appropriate presentation of technical information		☐ sufficient but not excessive technical information in body of report ☐ relevant and supportive technical information in appendices	Poor 1 2 Adequate 3 4 Excellent 5		

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
9.2 Other Aspects of Report	9.2.1 Main body of the report is of a reasonable length (25 to 40 pages)		☐ Yes	☐ less than 25 pages			
	9.2.2 Technical appendices are clearly identified and locations given		☐ Yes – clearly ☐ Yes – but not clearly enough ☐ No				
	9.2.3 Technical appendices are of high quality				Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		
	9.2.4 Data are presented fairly	Numbers are given Sources are documented		☐ numbers given ☐ sources documented	Poor 1 2 Adequate 3 4 Excellent 5		
	9.2.5 Effective use of tables and charts	Well presented Easy to read Fair	☐ No tables ☐ No charts/graphs ☐ Tables or charts/graphs not necessary or appropriate for this report	☐ effective tables ☐ effective charts/graphs	Poor 1 2 Adequate 3 4 Excellent 5 N/A 9		

Issues/ Requirements	Criteria	Considerations	General Checklist	Detailed Checklist	Rating8	Qualitative Assessment9	Other Comments
	9.2.6 Report is well-organized and easy to follow				Poor 1		
	9.2.7 Review is hindered by degree of Access to Information Act black-outs		☐ Yes – greatly ☐ Yes – slightly ☐ No				
10. Overall Assessment							
	10.1 Overall assessment				Poor 1 2 Adequate 3 4 Excellent 5		

Appendix B
Distribution of Reviewed Reports by Department/Agency

Distribution of Reviewed Reports by Department/Agency

Department/Agency	Number of Reports
Agriculture and Agri-Food Canada	3
Atlantic Canada Opportunities Agency	1
Communications Canada/Canada Information Office	3
Canadian Centre for Management Development	1
Canadian Centre for Occupational Health and Safety	1
Canada Customs and Revenue Agency	2
Canadian Space Agency	1
Canadian Heritage	11
Citizenship and Immigration Canada	1
Canadian International Development Agency	4
Canadian Institutes for Health Research	2
Correctional Services Canada	3
Foreign Affairs and International Trade	3
Fisheries and Oceans Canada	2
National Defense	2
National Defence/Veterans Affairs Canada	1
Economic Development Agency of Canada for the Regions of Quebec	3
Finance Canada	1
Health Canada	6
Human Resources Development Canada	5
Industry Canada	12
Indian and Northern Affairs Canada	10
Justice Canada	4
National Parole Board	1
National Research Council of Canada	5
Natural Resources Canada	10
Natural Sciences and Engineering Research Council	1
Office of Critical Infrastructure Protection and Emergency Preparedness	1
Public Service Commission	1
Royal Canadian Mounted Police	1
Status of Women Canada	1
Treasury Board Secretariat	2
Transport Canada	3
Veterans Affairs Canada	2
Western Economic Diversification	5
Total	115